



RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach

Pan, Xiaoyong; Shen, Hong Bin

Published in:
BMC Bioinformatics

DOI:
[10.1186/s12859-017-1561-8](https://doi.org/10.1186/s12859-017-1561-8)

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Document license:
[CC BY](#)

Citation for published version (APA):
Pan, X., & Shen, H. B. (2017). RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics*, 18(1), [136]. <https://doi.org/10.1186/s12859-017-1561-8>

RESEARCH ARTICLE

Open Access



RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach

Xiaoyong Pan^{1*} and Hong-Bin Shen^{2*}

Abstract

Background: RNAs play key roles in cells through the interactions with proteins known as the RNA-binding proteins (RBP) and their binding motifs enable crucial understanding of the post-transcriptional regulation of RNAs. How the RBPs correctly recognize the target RNAs and why they bind specific positions is still far from clear. Machine learning-based algorithms are widely acknowledged to be capable of speeding up this process. Although many automatic tools have been developed to predict the RNA-protein binding sites from the rapidly growing multi-resource data, e.g. sequence, structure, their domain specific features and formats have posed significant computational challenges. One of current difficulties is that the cross-source shared common knowledge is at a higher abstraction level beyond the observed data, resulting in a low efficiency of direct integration of observed data across domains. The other difficulty is how to interpret the prediction results. Existing approaches tend to terminate after outputting the potential discrete binding sites on the sequences, but how to assemble them into the meaningful binding motifs is a topic worth of further investigation.

Results: In viewing of these challenges, we propose a deep learning-based framework (iDeep) by using a novel hybrid convolutional neural network and deep belief network to predict the RBP interaction sites and motifs on RNAs. This new protocol is featured by transforming the original observed data into a high-level abstraction feature space using multiple layers of learning blocks, where the shared representations across different domains are integrated. To validate our iDeep method, we performed experiments on 31 large-scale CLIP-seq datasets, and our results show that by integrating multiple sources of data, the average AUC can be improved by 8% compared to the best single-source-based predictor; and through cross-domain knowledge integration at an abstraction level, it outperforms the state-of-the-art predictors by 6%. Besides the overall enhanced prediction performance, the convolutional neural network module embedded in iDeep is also able to automatically capture the interpretable binding motifs for RBPs. Large-scale experiments demonstrate that these mined binding motifs agree well with the experimentally verified results, suggesting iDeep is a promising approach in the real-world applications.

Conclusion: The iDeep framework not only can achieve promising performance than the state-of-the-art predictors, but also easily capture interpretable binding motifs. iDeep is available at <http://www.csbio.sjtu.edu.cn/bioinf/iDeep>

Keywords: RNA-binding protein, CLIP-seq, Deep belief network, Convolutional neural network, Multimodal deep learning

*Correspondence: xypan172436@gmail.com; hbshen@sjtu.edu.cn

¹Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Copenhagen, Denmark

²Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China

Background

RNA-protein interactions are involved in many biological processes, such as gene regulation and splicing [1]. Discovering the RNA-protein interactions has a great potential for further understanding the mechanisms behind those biological processes. For example, Argonaute (AGO) protein belongs to components of the RNA-induced silencing complex (RISC), which transfers microRNAs (miRNAs) to be bound with 3'UTR of their target genes, thereby result in mRNA degradation or translation repression [2]. Sequence-specific associations between RBPs and their RNA targets are mediated by binding domains, which recognize binding sites on RNAs. Where the RNA-protein binding sites on the RNAs are usually short sequences with 4 to 30 nucleotides long, typically referred as binding motifs. Detecting them can facilitate the deeper insights into post-transcriptional regulation.

Although there are many genome-wide RNA-binding protein detection techniques, such as RNAcompete [3], PAR-CLIP [4], they are still cost-heavy and time-intensive. Fortunately, with the advent of these high-throughput techniques, many useful genome-wide data associated with RBPs are generated rapidly, including specific binding positions on RNAs with proteins. These data provides important bases for developing computational approaches to predict the RBP binding sites by using the advanced computational methods [5–9].

At the very beginning of the methodology development of this field, predictors are mainly constructed by only using the sequence information. For instance, MatrixREDUCE simply fits a statistical mechanical model to infer the sequence-specific binding sites for transcription factors from sequences [10]. DRIMust discovers motifs by integrating the minimum hyper-geometric statistical framework with suffix trees for fast enumerating motifs [11].

Besides the high-throughput sequences, actually multiple sources of data are available from the genome-wide RNA-protein CLIP-seq data, such as sequences, structures, genomic context. Each source of data has a different kind of representation and correlation structure. A popular straightforward idea is to integrate these data to construct a predictor, which is expected to be very useful for enhancing the prediction accuracy. Two integration schemes have been widely used in the literatures:

1. Feature-level fusion. This type of fusion strategy is to encode the different sources into feature vectors, which will be concatenated together. For instance, the OliMoSS model has integrated tetranucleotide sequence, binding motifs and secondary structures to predict protein specific interactions on RNAs by simply concatenating the different sources of features

into one high-dimensional features (525-D) [12], which may result in difficulties for the following statistical learning process. For instance, the learning algorithm used in the OliMoSS is support vector machine (SVM), which will easily suffer from the curse of dimension problem. Similar strategy is also applied in DNA-protein binding sites prediction [13]. The other implementation of feature-level fusion is the multiple-kernel learning, which design multiple kernels for different features, and then combine them together [14, 15]. Similarly, GraphProt encodes the sequence and structure information to graph kernel to predict binding reference of RBPs [6].

2. Decision-level fusion. To solve the high-dimension space learning problem, decision level-based fusion system has been proposed. For instance, the iONMF [5] is a predictor for predicting RNA-protein interaction sites. It has trained a model for each of available resource data, e.g. kmer sequence, secondary structure, CLIP co-binding, Gene Ontology (GO) information, and region type. These independent 5 models will work independently, which have no interconnections between them during the training processes. The final prediction outputs of the whole system are the fusion of 5 independent predictions.

Despite the progresses of previously proposed methods, they have a shared drawback that the models were constructed on the features extracted from the observed data, where the frequent noise may make the subsequent classifiers learn wrong knowledge. Deep learning [16, 17] is a recently developed approach, which works in a hybrid multiple-layer abstraction way by mapping the observed data to a much high-level abstraction space, where the prediction model will be constructed. This new type of approach has provided much attractive solutions for integrating heterogeneous data and are effective in automatically learning complex patterns from multiple simple raw inputs.

One typical deep learning framework is known as the convolutional neural network (CNN) [18]. The advantage of CNN is that it does not separate feature extraction and model learning into two independent steps any more as done in the traditional statistical learning algorithms. Instead it simultaneously learns features and classification models from the original input in a data-driven way, which will reduce the potential mismatch effects between the feature extraction and learning classification models. The CNN model has been applied in the binding proteins prediction of DNA or RNA. For instance, a recent CNN-based deep learning approach DeepBind was proposed to predict sequence specificities for protein binding RNA/DNA [8]. Similarly, the DeepSEA [19] utilized the deep CNNs to learn regulatory sequence motifs for

predicting DNA functions from chromatin profiling data; Basset [20] trained analogous deep CNN models to learn impacts of DNA sequences variants on chromatin regulation from large-scale DNase-seq data. These studies have shown that the convolution operation in CNN is able to scan a set of weight matrix (filters) across the input sequences to recognize relevant patterns that respond to motifs, like patterns corresponding to edges and curve fragments in images [21, 22], resulting in better prediction accuracies [5, 12].

The deep belief network (DBN) is another deep learning algorithm to learn high-level features from large-scale data [23], which is also a recent popular choice for constructing the computational models. For example, the deepnet-rbp fuses the structural and k-mer sequence features to predict RBP interaction sites [24] using DBNs. DANN trains a DBN to annotate non-coding variants [25], which is able to capture non-linear abstraction features. We also developed a model called IPMiner by applying the stacked autoencoder to learn high-level features for predicting RNA-protein interactions from raw sequence composition features, and it yielded promising performance compared to other sequence-based methods [26]. It's worth noting that many studies have shown that the CNN and DBN hold their own advantages due to different deep learning architectures, e.g. CNN is more appropriate for sequence data and DBN prefers to the numeric inputs. This motivates us to consider how to integrate the merits of CNN and DBN for better prediction of RBP binding sites and find the sequence motifs.

In this study, we propose a multimodal deep learning framework iDeep, a hybrid framework with CNNs and DBNs, to better integrate multiple heterogeneous data sources for predicting RBP interaction sites on RNAs (Fig. 1). For the data represented by the binary or numeric features, the DBN networks will be used; While for the sequence data, the CNN network will be applied. Different deep network models will be trained and tuned together from the top shared layer to the individual bottom layers using backpropagation, and then the shared latent features are captured across them. Compared to the existing approaches, the iDeep has the following merits: 1) the iDeep is constructed with a deep learning structure, and it consists of multiple neural networks stacked together [16, 17], where the outputs of each layer are the inputs of successive layer. Such layer-by-layer learning helps to reduce the noise effects in the original input. 2) The iDeep successfully integrates the CNN and DBN for dealing with the different sources of protein-RNA binding related data to enhance the discrimination ability. The CNN is able to capture regulatory motifs, which are recurring patterns in RNA sequences with a biological function. The DBN learns high-level features regarded as a joint distribution determined by hidden variables

for different inputs. 3) The hybrid framework of flexible multimodal learning and fusion at an abstraction level makes the iDeep handle different features in an easy manner. The top shared hidden layer at the fusion level will help discover the shared properties across different modalities [27, 28].

Results

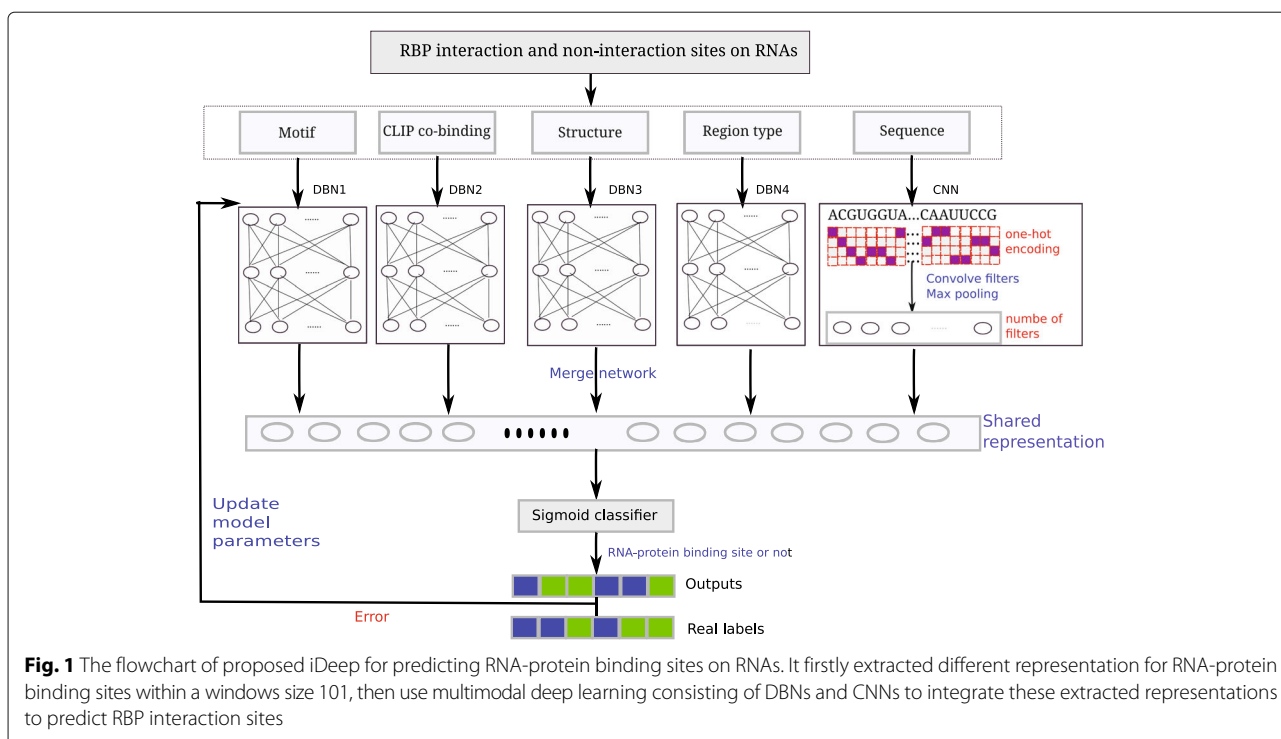
In this study, we evaluated iDeep on independent testing datasets, and also compared it with the performance of DBN and CNN from individual sources of data. To demonstrate the advantage of iDeep, some state-of-art predictors of iONMF, DeepBind, and Oli were also compared. Besides, a large-scale analysis has been conducted to demonstrate the discovered binding motifs using iDeep.

The iDeep's performance

To demonstrate the ability of iDeep for predicting RNA-protein binding sites, we evaluate iDeep on independent testing sites (see the dataset section). We firstly use 4000 training samples for model training, 1000 validation sites are evaluated at the end of each training epoch to monitor the convergence. For each experiment, iDeep is trained with the same initializations. After we obtain the trained model, we apply it to predict binding sites for 1000 independent testing samples. The ROC on 31 experiments are shown in Fig. 2. It indicates that iDeep yields different performance on different experiments with huge margin, the AUC ranges from 0.68 for protein ELAVL1-MNase to 0.98 for protein PUM2. In addition, iDeep achieves the AUC greater than 0.90 on 23 of 31 experiments, and the average AUC of iDeep on all experiments is 0.90. To make a more accurate performance estimate of iDeep, we also run the 5-fold cross-validation to evaluate the performance, iDeep yields the average AUC 0.91 (Additional file 1: Figure S1). The results indicate that iDeep accurately predict RBP binding sites on a genome-wide scale.

Comparing iDeep with other state-of-the-art methods

We firstly compare it with state-of-the-art method iONMF, which has shown better performance than other existing methods [5], such as GraphProt [6] and RNA-Context [29]. As shown in Table 1, we can see that iDeep outperform iONMF on most of the 31 experiments, the average AUC of the 31 experiments increases from 0.85 ± 0.08 of iONMF to 0.90 ± 0.08 of iDeep. Furthermore, for some experiments, it improves the AUC over 24%, such as for protein hnRNPL-2, the AUC increases from 0.66 of iONMF to 0.82 of iDeep. In addition, iDeep also performs better than other matrix factorization-based methods NMF [30], SNMF [31] and QNO [32], which achieves the average AUC of 0.83 ± 0.10 , 0.71 ± 0.14 , 0.79 ± 0.12 on 31 experiments, respectively.



We further compare iDeep with another protein-specific method Oli [12], which yields an average AUC of 0.77 ± 0.16 , and 17% lower than the iDeep. We find that it has a bigger performance variance than other tested methods. For example, Oli performs very bad on some experiments, e.g. AUC 0.39 on hnRNPL-1 protein, but on some experiments, its performance is very good, e.g. 0.94 on PUM2 protein. For the DeepBind [8] approach using the same parameters of CNN integrated in iDeep, it achieves an average AUC 0.83 ± 0.12 across 31 experiments, which performs worse than iDeep. The reason is that DeepBind cannot yield promising performance across all 31 experiments from only sequences.

To demonstrate the merits of the designed framework of iDeep, we also compare iDeep with its own variant iDeep-kmer, whose input modalities are kmer, region type, clip-cobinding and structure using the same network architecture. The only difference is that iDeep uses CNN sequence and motif modalities instead of high-dimensional kmer modality. As indicated in Table 1, iDeep-kmer yields an average AUC of 0.87 ± 0.09 , which is worse than iDeep, indicating that CNN sequence and motif modality have better discriminating ability than high-dimensional kmer modality. On the other hand, iDeep performs faster than iDeep-kmer both in the training and testing steps.

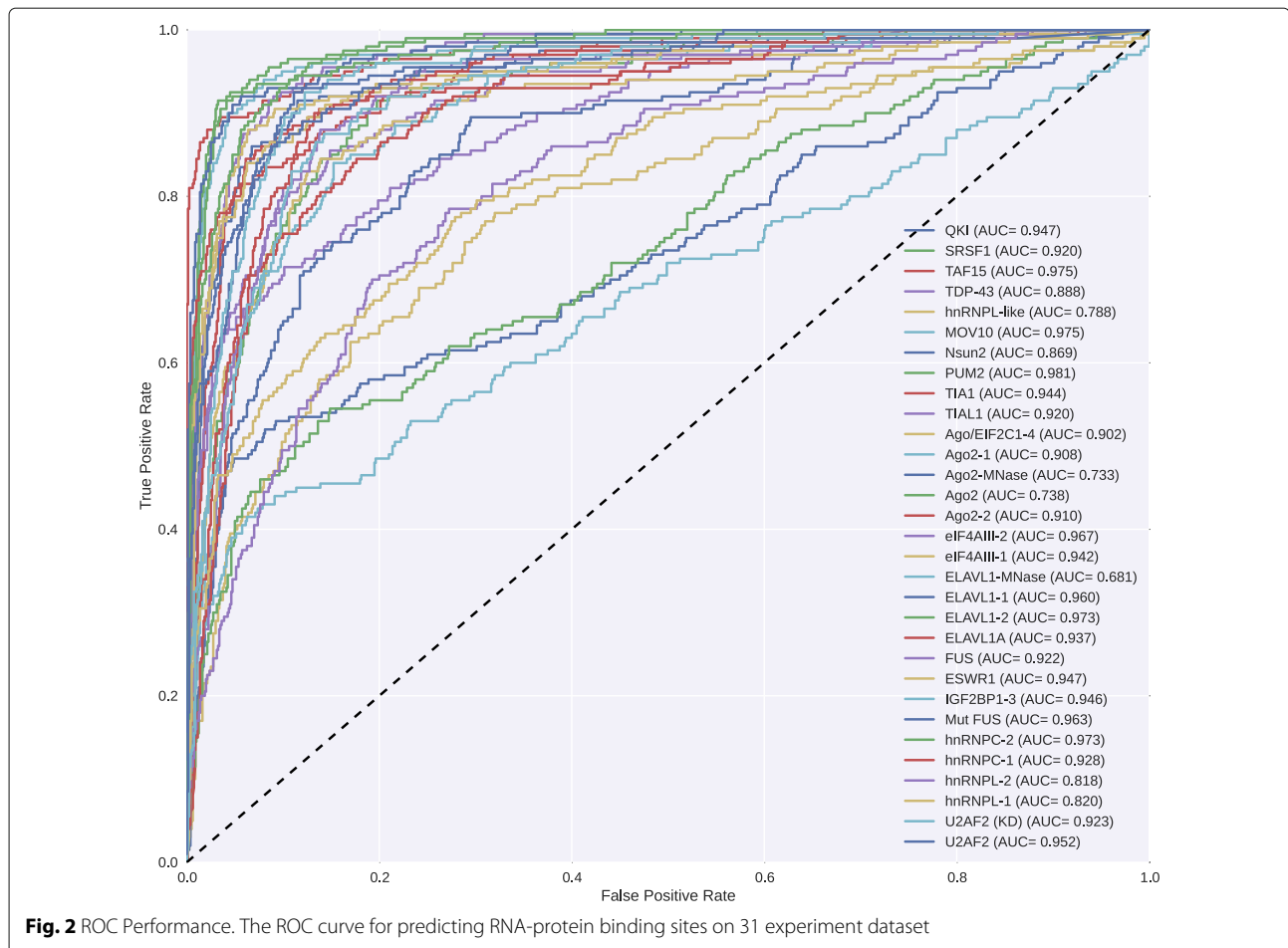
Overall, compared to other 6 tested methods, iDeep yields the best performance on 18 of 31 experiments and the same AUCs on other 8 experiments. And it achieves

a little lower AUC only on 5 of the 31 experiments, but it still yields the AUCs above 0.90 on 4 of them. For those experiments with AUCs below 0.90 in other six methods, iDeep's performance is very encouraging. These results indicate that iDeep's promising performance.

Comparison between individual modalities

To show the advantage of integrating multiple modalities of data, we also tested the performance on individual modalities. The average AUCs of 31 experiments for region type, clip-cobinding, structure, motif and CNN sequence are 0.73 ± 0.11 , 0.74 ± 0.11 , 0.71 ± 0.12 , 0.71 ± 0.08 and 0.83 ± 0.12 , respectively, indicating that individual deep networks have the ability of learning high-level features for RBP binding sites prediction. From the results, we can see that CNN sequence modality yield the best average performance with roughly 12% improvement over the second most informative region type. And CNN sequence yields higher AUC on 22 experiments due to sequence specificities of binding RNA [8], where CNN sequence can automatically learn binding motifs as feature representations for subsequent classifications. The other 4 modalities achieves similar average AUCs on all experiments without a big difference. Furthermore, we also test the performance of DBN with only kmer modality, it yields the average AUC of 0.76 ± 0.13 on 31 experiments, which is found much worse than CNN sequence modality.

As indicated in Fig. 3, there exists big performance differences on individual experiments for different



modalities. For instance, on U2AF2 (KD) experiment, the 5 individual modalities achieve the AUC of 0.66, 0.65, 0.53, 0.72 and 0.91, respectively. The CNN sequence modality obtains AUC 0.91, outperforming other 4 modalities. While for experiment ELAVL1-MNase, they yield the AUCs of 0.67, 0.70, 0.67, 0.54, and 0.54, respectively. The CNN sequence achieves the worst AUC of 0.54 and the clip-cobinding modality has the best AUC of 0.70. The results showed that there were huge differences between different modalities on different experiments.

Among the 5 Ago2 experiments, structure modality performs a little better on 3 of them. It is because that Ago2 protein requires specific RNA structure binding interfaces [33]. The motif and CNN sequence modalities perform worse than other modalities on the 5 Ago2 experiments. The reason is that Ago2 protein has a PAZ domain and a PIWI domain, but there are no related binding motifs for them in CISBP-RNA database [34], and hence deep network of motif and CNN sequence modalities cannot learn high discriminating features for predicting Ago2 binding sites on RNAs. Although motif and CNN sequence modality are not able to detect binding sites for Ago2

with high accuracy, other modalities can complement with them. The more diversity different modalities have, the more accurate the integrated method is [35]. So integrating the 5 different modalities using multimodal deep learning makes iDeep perform much better than individual modalities.

Based on the above results, we can have the following conclusions: (1) No single modality can beat others on all datasets, their performance varies on different datasets. (2) The deep network (CNN and DBN) of input modalities are able to learn high-level features with stronger discriminating ability for RBP interaction sites. (3) Integrated iDeep performs better than deep networks of individual modalities, it is because that multimodal deep learning is able to learn shared representation across multiple modalities with strong discriminating ability for RNA-protein binding sites.

The correlations between different modalities in deep architecture

In the proposed iDeep model, we integrated 5 sources of data for an ensemble prediction. It will be interesting to see how the 5 independent modalities will complement

Table 1 The AUC performance comparison between iDeep and other methods on 31 experiments

Protein	iDeep	iONMF	NMF	SNMF	QNO	Oli	iDeep-kmer	DeepBind
1 Ago/EIF	0.90	0.89	0.89	0.85	0.87	0.61	0.87	0.69
2 Ago2-MNase	0.73	0.71	0.69	0.66	0.69	0.51	0.67	0.53
3 Ago2-1	0.91	0.81	0.81	0.76	0.83	0.80	0.82	0.81
4 Ago2-2	0.91	0.84	0.82	0.79	0.82	0.80	0.83	0.81
5 Ago2	0.74	0.73	0.71	0.65	0.66	0.53	0.65	0.58
6 eIF4AIII-1	0.94	0.92	0.91	0.78	0.95	0.92	0.95	0.93
7 eIF4AIII-2	0.97	0.93	0.93	0.67	0.64	0.93	0.94	0.93
8 ELAVL1-1	0.96	0.91	0.89	0.71	0.80	0.89	0.95	0.90
9 ELAVL1-MNase	0.68	0.71	0.70	0.68	0.70	0.49	0.66	0.54
10 ELAVL1A	0.94	0.94	0.93	0.91	0.92	0.84	0.95	0.87
11 ELAVL1-2	0.97	0.95	0.94	0.90	0.95	0.88	0.97	0.91
12 ESWR1	0.95	0.87	0.85	0.80	0.85	0.81	0.92	0.88
13 FUS	0.92	0.81	0.73	0.55	0.65	0.85	0.87	0.92
14 Mut FUS	0.97	0.96	0.95	0.91	0.94	0.82	0.97	0.91
15 IGFBP1-3	0.95	0.93	0.92	0.89	0.91	0.57	0.93	0.68
16 hnRNPC-1	0.93	0.95	0.93	0.45	0.63	0.88	0.92	0.95
17 hnRNPC-2	0.97	0.97	0.96	0.48	0.70	0.94	0.95	0.97
18 hnRNPL-1	0.82	0.74	0.73	0.70	0.77	0.39	0.79	0.76
19 hnRNPL-2	0.82	0.66	0.62	0.56	0.61	0.47	0.72	0.74
20 hnRNPL-like	0.79	0.69	0.67	0.63	0.68	0.56	0.70	0.70
21 MOV10	0.97	0.96	0.96	0.89	0.92	0.78	0.97	0.80
22 Nsun2	0.87	0.81	0.80	0.69	0.82	0.75	0.81	0.84
23 PUM2	0.98	0.93	0.92	0.86	0.89	0.94	0.98	0.93
24 QKI	0.95	0.84	0.77	0.52	0.62	0.92	0.92	0.95
25 SRSF1	0.92	0.85	0.85	0.73	0.86	0.84	0.85	0.85
26 TAF15	0.97	0.91	0.89	0.82	0.91	0.80	0.95	0.95
27 TDP-43	0.89	0.84	0.78	0.45	0.57	0.88	0.85	0.89
28 TIA1	0.94	0.93	0.92	0.86	0.90	0.84	0.96	0.90
29 TIAL1	0.92	0.87	0.86	0.73	0.85	0.83	0.90	0.87
30 U2AF2	0.95	0.82	0.74	0.61	0.70	0.86	0.91	0.95
31 U2AF2(KD)	0.92	0.80	0.74	0.60	0.74	0.84	0.88	0.91
Mean	0.90±0.08	0.85±0.08	0.83±0.10	0.71±0.14	0.79±0.12	0.77±0.16	0.87±0.09	0.83 ± 0.12

The performance of iONMF, NMF, SNMF and QNO are taken from [5]. DeepBind, Oli and iDeep-kmer perform on the same data with iDeep, and iDeep-kmer used kmer to replace CNN sequence and motif modalities in iDeep

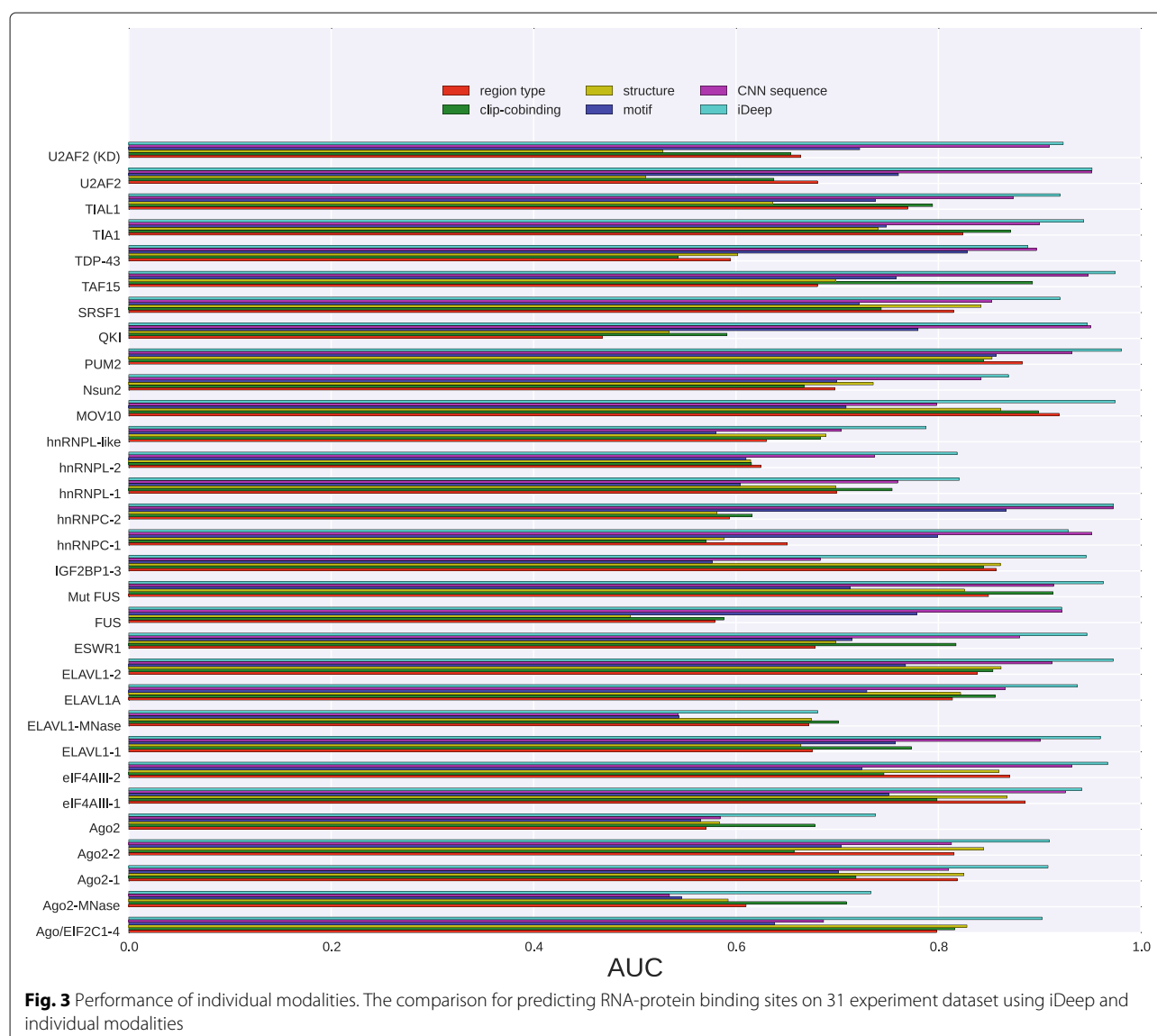
The boldface indicates this performance is the best among the compared methods

with each other. We thus investigate the pairwise correlation between the different modalities region type, clip-cobinding, structure, motif, CNN sequence across 31 experiments. In addition, we also demonstrate the correlations between the 5 modalities and unintegrated high-dimensional k-mer modality.

We calculate the Pearson correlation coefficients (PCC) based on the AUCs of 31 experiments from individual modalities. If two modalities have high PCC, they perform similarly across all 31 experiments. As illustrated in Fig. 4, there are two obvious subgroups between the 6

modalities. The region type, clip-cobinding and structure formed the first group; kmer, motif and CNN sequence formed the other group. These results show that different modalities contain various signals, and they can complement with each other via integration in iDeep.

The region type and structure modalities have a PCC of 0.89, showing that they are highly correlated. It is because the same region type may have similar structures, they share redundant information for predicting binding sites. CNN sequence and kmer also have very high PCC of 0.97, indicating that they are also highly related. As



demonstrated in the iONMF [5], kmer modality can capture binding motifs. CNN sequence also learns motifs using CNN network [8, 19]. In addition, both of them are highly correlated to motif modality with PCCs of 0.91 and 0.92, respectively. It indicates that the high-level features learned from CNN sequences and kmer are closely related to binding motifs, which is consistent with previous findings. In summary, both the modalities try to learn binding motifs, so they share similar signals associated with motifs for RBPs across the 31 experiments. That is also the reason why we used CNN sequence instead of high-dimensional kmer in iDeep.

The iDeep is able to discover new binding motifs

The iDeep can predict RBP binding sites on RNAs with high accuracy, however the principles behind it are still

not easily interpretable. So here we further use iDeep to discover binding motifs for RBPs. In previous methods [5, 12], they focus on directly detecting nucleotide binding sites on RNAs from extracted features, but did not introduce the motifs during feature learning. Although iONMF tries to infer the binding motifs after model training, it totally depends on the input kmer sequences and defines a background distribution. In addition, it limits the learned motifs to size k , which requires optimization for searching potential motifs and the time cost exponentially increases with k .

To explore the learned motifs, we investigate the convolve filters of the convolutional layers from CNN module in iDeep, and convert them into position weight matrices (PWM), which is matched against input sequences to discover binding motifs, like DeepBind [8] and Basset

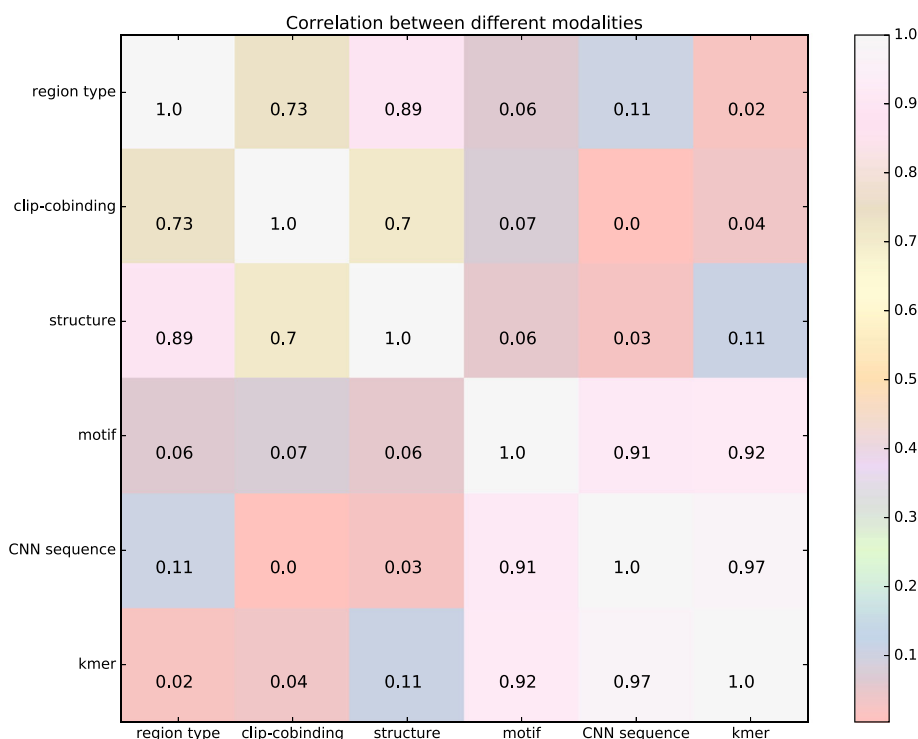


Fig. 4 The correlation between different modalities on 31 experiment dataset. The pearson correlation coefficient values are calculated using the AUCs from 31 experiments for individual modalities

[20] (Additional file 2). Then, these discovered motifs are aligned against 102 known motifs in study [34] from CISBP-RNA using the TOMTOM algorithm [36]. In addition, we also provide enrichment and occurrences analysis for the predictive motifs (Additional file 2).

Using p -value < 0.05 , iDeep captures most of informative motifs for individual proteins. The significantly matched known motifs for individual experiments are listed in Fig. 5, where 15 experiments with known motifs in study [34] are included. As can be seen from Fig. 5 that the iDeep is able to mine known motifs for 13 of 15 experiments. For example, there are 5 known motifs (M031, M108, M112, M127, M232) in study [34] for protein ELAVL1-2, and all of them have been correctly discovered by iDeep. Fig. 6a illustrates the heatmap of learned weights of convolve filters of CNN and corresponding matched known motifs for these filters for protein TDP-43, IGFBP1-3, and Ago2, respectively, and their enrichment adjusted p -value are $1.12e^{-15}$, $2.38e^{-3}$ and $2.38e^{-15}$, respectively. Besides the already well-known motifs discovered by iDeep, it is able to find some novel motifs. For instance, for protein TDP-43, currently there are no verified motifs for it in CISBP-RNA database, although TDP-43 have been discovered to bind to thousands of RNAs in neuron [37]. Figure 6b shows the hierarchical

clustering of 102 new filters (motifs) for protein TDP-43 discovered by iDeep. Of them, two newly identified motifs with adjusted p -value $1.90e^{-15}$ and $3.35e^{-8}$ for protein TDP-43 are illustrated in Fig. 6c. These new motifs (Additional file 3: Figure S2) will provide important clues for further wet-lab verifications. Due to the incompleteness of current database or the insensitivity of the TOMTOM software, many mined motifs by iDeep cannot be matched. It is expected that with more RBP motifs being verified and deposited in the database, more matches will be found. All discovered motifs by iDeep and their motif enrichment scores are available at https://github.com/xypan1232/iDeep/tree/master/predicted_motifs.

Discussion

In iDeep, we do not integrate high-dimensional k-mer and GO features, which possibly causes the over-fitting problem when calculating the partition functions. In addition, for other 5 integrated features in iDeep, dropout layer was applied for both CNN and DBN. It randomly sets 0s for some unit activations with certain probabilities, which can avoid over-fitting for model training [38]. Besides, we also introduce batch normalization and early stopping to prevent the possible over-fitting during the training process [39].

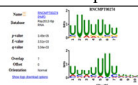
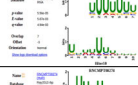
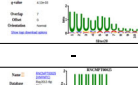
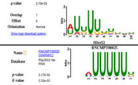
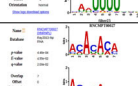
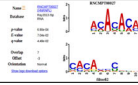
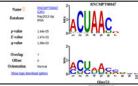
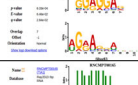
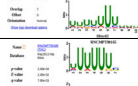
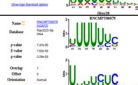
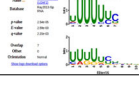

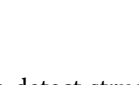
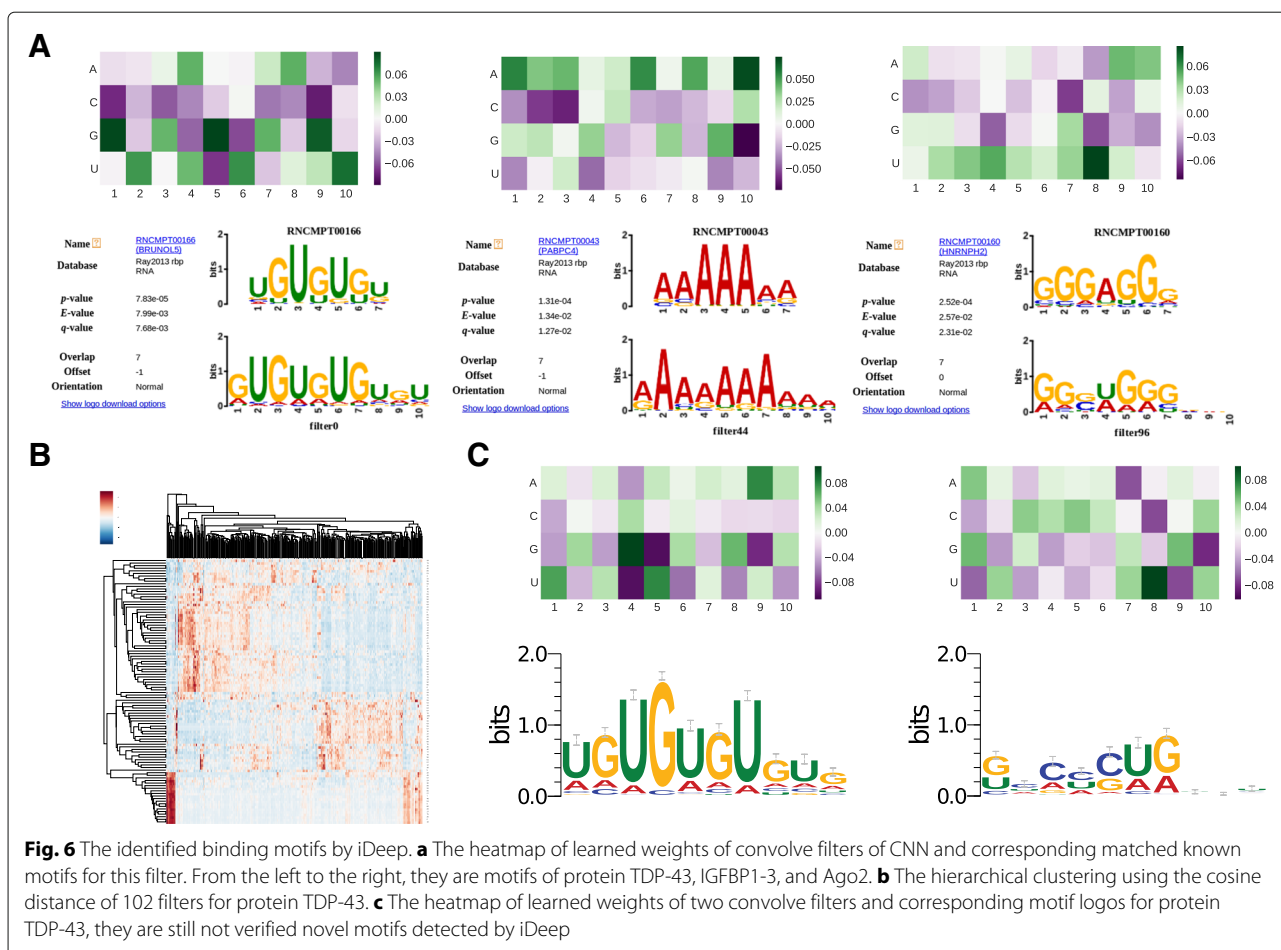
Protein	Known motifs	Matched predicted motifs	One example
8 ELAVL1-1	M031,M108,M112,M127,M232	M031,M108,M112,M232	
9 ELAVL1-MNase	M031,M108,M112,M127,M232	M031,M108,M112,M127,M232	
11 ELAVL1-2	M031,M108,M112,M127,M232	M031,M108,M112,M127,M232	
15 IGFBP1-3	M032,M163	-	-
16 hnRNPC-1	M025	M025	
17 hnRNPC-2	M025	M025	
18 hnRNPL-1	M027,M089	M027	
19 hnRNPL-2	M027,M089	M027	
20 hnRNPL-like	M1696	-	-
24 QK1	M046	M046	
25 SRSF1	M102,M103,M104 M105,M106,M154	M102,M103,M104 M105,M106,M154	
28 TIA1	M075,M156	M075,M156	
29 TIAL1	M075,M156	M075,M156	
30 U2AF2	M077	M077	
31 U2AF2(KD)	M077	M077	

Fig. 5 iDeep captures known motifs in [34] from CISBP-RNA for proteins. We only compared our predicted motifs against known motifs in study [34] and the motif name is from CISBP-RNA. If there is no motifs for this protein, then we ignore them. - means no matched motifs in our predictions with e-value cut-off 0.05

In our 5 modalities integrated in iDeep, CNN sequence modality outperforms other modalities on most experiments. But for some proteins, such as Ago2, it performs worse than structure modality, indicating structure information has better informative signals for Ago2 binding sites. Currently we just use simple probabilities predicted from RNApfold [40] as the input features, which contain some noises due to the accuracy below 100%. So in future work, we will extend the CNN to structures, and design CNN to find high-level structure motifs for RBP binding sites. As done in GraphProt [6], they

apply graph encoding to detect structure motifs. We can adopt similar strategy for encoding RNA structure to 6 elements (stem, multiloop, hairpin loop, internal loop, bulge and external regions), which can be fed into CNN for learning structure motifs automatically to further improve iDeep's performance. In addition, Ago2 binding specificity is provided primarily by miRNAs [2], the expressed miRNAs in a given cell type greatly influences Ago2-RNA interactions, resulting in a much more variable and cell type-dependent binding motifs than RNA-binding proteins which bind their mRNA targets directly.



Integration of miRNA expression as an additional modality will conceivably improve the accuracy of iDeep for Ago2 proteins.

The iDeep outperforms other state-of-the-art methods with the average AUC of 0.90 on 31 experiments. iDeep goes a step further besides predicting the RBP binding sites, it also is able to identify the high-level motifs. The motifs help to understand the correlations of the binding sites and their biological functions. In addition, iDeep also discovers some novel binding motifs besides those reported motifs in CISBP-RNA, we expect to verify those novel motifs by investigating whether the genes with the same predicted motifs are significantly associated with certain functions. Meanwhile, these candidate motifs could provide a quick guide for the wet-lab experiments to avoid very time-consuming search.

In addition, when selecting sequences in windows for creating sequence motifs using WebLogo [41] for each convolve filter, we use the cutoff 0.5 of the maximum activation of this filter over all sequences (greater than 0). This cutoff will remove some positions with no significant information to be selected for generating motifs.

However, if we use other cutoffs, e.g. 0, then it might introduce noises. Thus, ReLU may result in the information loss, but it has been found helpful in avoiding the vanishing gradient problem and learning non-linear features [42].

Despite the promising performance of iDeep, there are still promising avenues to explore the ability of deep learning. Currently we only use the standard CNNs for sequences and similar DBNs for other data modalities with only different number of hidden neurons, which should be designed specifically for different input data. Besides, more advanced network architecture could be designed according to the special characteristics of different input data. For example, DanQ designed a hybrid convolutional and recurrent neural network to predict the functions from non-coding DNA sequences [43]. It uses CNN to detect regulatory motifs from sequences, followed by bi-directional recurrent layer to capture long-term dependencies between motifs. Furthermore, instead of learning high-level features using deep learning, another study aims at automatically learning hand-designed optimization algorithms, which can

exploit the structures in network architecture of interest [44]. All these studies indicate that we can further improve the structure of current iDeep to improve the performance in the future.

Conclusion

In this study, we present a deep learning based hybrid framework to integrate different sources of data to predict RNA-protein binding sites on RNAs from CLIP-seq data, which yields promising performance on large-scale experiment data. The iDeep has the following advantages: (1) It trains deep neural network on individual sources of data to learn high-level representations for predicting RNA-protein interaction sites. (2) Different from other black-box machine learning based approaches, iDeep is able to discover the interpretable binding motifs, which provides better biological insights into RBPs compared to other black-box machine learning models. (3) It makes use of multimodal deep learning to extract shared features across different sources of data, with the hypothesis that no single one can overwhelm others across all datasets. Multimodal deep learning is able to better fuse them and achieve better performance on all datasets. Our proposed deep learning framework provides a powerful approach and choice for heterogeneous data integration.

Methods

In this section, we firstly introduce the CLIP-seq datasets and multiple features extracted in this study, then we design a multi data source driven multimodal deep learning framework to integrate them for predicting RNA-protein binding sites on RNAs.

Datasets

In this study, to compare with the existing state-of-the-art methods, we used the same benchmark human dataset as iONMF [5], which was downloaded from <https://github.com/mstrazar/ionmf>. In this dataset, the CLIP-seq data consists of 19 proteins with 31 experiments and the annotation is based on hg19. As described in the iONMF, each nucleotide within clusters of interaction sites derived from CLIP-seq were considered as binding sites. To reduce the redundancy, the positive binding sites were further randomly sampled with the highest cDNA count and without consecutive sites on genome. Finally, from those sites with less than 15 nucleotides apart, only one site with the highest cDNA counts was selected as the positive sample. The negative sites were sampled from genes that were not identified as interacting in any of 31 experiments. In the experiments, a total 4000 crosslinked sites are used for training purpose, 1000 samples for model optimization and validation, and the other 1000 samples for independent testing.

Feature encoding

Feature encoding is critical for a statistical machine learning model. In order to integrate the merits from both the sequence and numeric features, the iDeep model makes use of 5 different groups of features, i.e., sequence, structure, clip-cobinding, region type and motif features. A scale window of [-50, 50] centering the crosslinked sites is used to generate the feature vectors, which is the same as iONMF [5].

1. **Region type.** this feature value is assigned to each position within the window using one of the 5 types (exon, intron, 5'UTR, 3'UTR, CDS) from Ensembl annotation [45], resulting in $101 \times 5 = 505$ dimensional features.
2. **clip-cobinding.** This feature represents the correlation among 31 experiments. For each experiment, the cDNA counts at each position within the window was reported in the remaining 30 experiments, assign 0 for zero cDNA counts or 1 otherwise. We obtain 30 binary values for each nucleotide in the window, resulting in $101 \times 30 = 3030$ dimensional features.
3. **Structure.** RNAplfold in ViennaRNA Package 2.0 [40] with parameter -u 1 is used to calculate the probability of RNA secondary structure for each nucleotide within window, resulting in 101 dimensional features.
4. **Motif.** Motif scores are used for numerical representation of the RNA sequences [46]. We firstly downloaded 102 human RBP binding motifs from CISBP-RNA [34], then Cluster-Buster [47] was employed to score RNA sequences for binding sites clusters. For individual sequence, we can get a score per motif, resulting in a 102 dimensional features.
5. **CNN sequence.** The sequence is encoded into a 101×4 binary matrix corresponding to the presence of A, C, G, U, which is fed into CNN to obtain high-level sequence feature.

It's worth noting that since the iDeep model is constructed with the CNN algorithm, the 25856-D kmer and 39560-D GO features originally used in the iONMF are not used in our model. The main reasons are: 1) the GO features has been indicated of lower discriminating power than other sources of data [5] and 2) these two features are of too high dimensions, even more than the training samples, which easily leads to over-fitting and dimension disaster for neural networks. We also added two new feature encoding methods, which have not been applied in the iONMF, i.e., the sequence and motif features. Our results below will show that the new sequence feature encoding are critically important for CNNs to learn binding motifs, and the motif features

based on known motifs in CISBP-RNA database are useful to correlate with functional regulatory regions in RNA sequences.

Convolutional Neural Network

Convolutional Neural Network (CNN) is inspired by biological processes, it consists of one or more convolutional layers, followed by the max pooling layers. And it enforces a local connectivity pattern between neurons of layers to exploit spatially local structures. In this study, CNN is used to capture non-linear sequence features, e.g. motifs, and pull out some high-level features associated with RBP binding.

Here RNA sequence is one-hot representation encoded into a 101 x 4 binary matrix, whose columns correspond to A, C, G and U [8, 19]. Then the inputs are convolved with tunable patterns called filters, which are weight parameters corresponding to binding motifs and learned from RNA sequences. After convolution, a rectified linear ReLU is applied to avoid the vanishing gradient problem existing in deep learning research. Finally, a max pooling operation is used to pool adjacent positions within a small window, which can reduce the number of parameters and yield invariance to small sequence shifts.

Deep Belief Network

Deep Belief Network (DBN) consists of multiple layers of Restricted Boltzmann machines (RBMs) [48], which learns model parameters in bottom-up style and layer-wise, but it is only able to learn abstract structure from one input source of data.

Some of our extracted input features are binary, such as region type of nucleotides. RBM is developed for binary-valued inputs, which is a graphical model with visible $v \in \{0, 1\}$ and hidden units $h \in \{0, 1\}$. Its joint distribution of hidden and visible variables are defined as follows:

$$P(\mathbf{v}, \mathbf{h}, \theta) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}, \theta)) \quad (1)$$

where $E(\mathbf{v}, \mathbf{h}, \theta)$ is defined:

$$E(\mathbf{v}, \mathbf{h}, \theta) = - \sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i h_j w_{ij} \quad (2)$$

where v_i and a_i are binary state and bias for visible unit i , respectively. h_j and b_j are the binary states and bias for hidden unit j respectively.

The partition function Z is calculated by summing over hidden and visible variables, which is optimized using maximum likelihood estimation based on Contrastive

Divergence algorithm [17]. Besides, we also extract structure probability features, which are real-valued inputs, and its extension Gaussian RBMs are developed for modelling real-valued inputs [49]. The parameters weight matrix and biases are updated using a gradient descent algorithm [17].

DBN is comprised of multiple RBMs, Here we take a DBN with two hidden layers as example:

$$E(\mathbf{v}, \mathbf{h}, \theta) = -\mathbf{v}^T \mathbf{W}^{(1)} \mathbf{h}^{(1)} - \mathbf{h}^{(1)T} \mathbf{W}^{(2)} \mathbf{h}^{(2)} \quad (3)$$

where $h^{(1)}$ and $h^{(2)}$ are hidden units for two hidden layers, and $W^{(1)}$ and $W^{(2)}$ are weight parameters for visible-to-hidden and hidden-to-hidden connection.

DBN is able to capture high-level features from individual modalities, but it cannot interactively learn unified feature representations across them.

Multimodal deep learning for Predicting RNA-protein interaction sites

Considering the heterogeneous representations of RBP binding sites, multimodal deep learning is developed to learn shared features across different sources of data [27]. It consists of multiple layers of neural networks, which can automatically learn high-level features hidden in original features [16, 17] and achieve a huge success in different applications. In this study, we use CNNs and DBNs as the building blocks for deep learning framework shown in Fig. 1. It adds an additional layer to combine the outputs from multiple DBNs and CNNs for different inputs. During feature learning, individual DBNs and CNNs are pre-trained independently and concatenated together for final joint training using backpropagation. In each training epoch, it will automatically tune the learned parameters in respective models. After several training epochs, it learns shared representations across region type, clip-cobinding, structure, motif and CNN sequence for subsequent classification. In addition, it can also learn better features for individual modalities via backpropagation when multiple modalities exist.

We apply multimodal deep learning to integrate different sources of data to predict RNA-protein binding sites on RNAs. It first extracts different representations of different sources of data from CLIP-seq data, which are subsequently integrated using multimodal deep learning to predict RNA-protein binding sites. The flowchart is shown in Fig. 1.

In this study, we set the maximum number of epoch to 20, the batch size is 100. The neural network models are optimized using RMSprop algorithm [50] to learn all model parameters, including those convolution filters of CNNs. Validation dataset is evaluated to monitor the convergence during each epoch of the training process, so the training process can be stopped early.

Table 2 The number of neurons used in fully connected layer (FCL) for each DBN

Feature	# of neurons of first FCL	# of neurons of second FCL
CLIP-cobinding	768	256
Structure	128	64
Region type	256	128
Motif	128	64

The iDeep is implemented in python using keras 1.0.4 library <https://github.com/fchollet/keras>. The model architecture consists of hybrid CNNs and DBNs for individual inputs and additional layer for merging them.

For sequence modality, its one-hot encoding is fed into CNN to learn high-level motif features. The parameter nb_filter (number of motifs) is 102. The size of the significantly verified RBP binding motifs in CISBP-RNA database is 7 and the 7-mer motifs can consistently score well [34]. According to the suggestion by DeepBind that the parameter filter_length (motif width) should be 1.5 times the verified motif size, thus we set filter_length = 10.

The architecture of DBN for input modalities clip-cobinding, Structure, Region type and Motif consists of fully connected layer and dropout layers (Additional file 1). In iDeep, for each DBN from individual modalities, we configure different number of hidden units for two Fully connected layer (FCL) listed in Table 2, and the dropout probability for each dropout layer is 0.5. To evaluate the performance of predicting RBP binding sites, we use Receiver Operating Characteristic(ROC) curve and calculate the area under the ROC curve (AUC).

Baseline methods

There are many computational methods developed for predicting RNA-protein binding sites. such as iONMF, Oli, DeepBind, GraphProt and RNAContext. As indicated in [5], iONMF performs a little better than GraphProt, and much better than RNAContext. In [12]), Oli with only tetranucleotide frequency features yield better performance than its variant OliMoSS for predicting RBP binding sites. So in this study, we compared iDeep with other state-of-the-art iONMF, DeepBind and Oli. iONMF integrates multiple data using orthogonality-regularized nonnegative matrix factorization, it discovers the hidden modules from non-overlapping features for RNA-protein interactions. Oli applied linear SVC to classify protein-RNA binding sites based on their extracted tetranucleotide frequency features. To compare with Oli fairly, grid-search was used to select the best parameter for linear SVC of Oli in individual experiments, and the implementation from scikit-learn package was used in this study [51]. For DeepBind, it only uses CNN from sequences to predict RBP binding sites.

Additional files

Additional file 1: Figure S1. The AUCs of 5-fold cross-validation across 31 experiment datasets using iDeep. (PDF 24 kb)

Additional file 2: Supplementary text and Table. Some details of iDeep. The principles about how to identify binding motifs by iDeep, the architecture of deep belief network and the discovered number of known motifs in CISBP-RNA. (PDF 127 kb)

Additional file 3: Figure S2. The novel motifs still not verified by other studies are discovered by iDeep. (PDF 65 kb)

Abbreviations

AUC: The area under the ROC curve; CNN: Convolutional neural network; DBN: Deep belief network; FCL: Fully connected layer; GO: Gene ontology; PCC: Pearson correlation coefficient; PWM: Position weight matrices; RBPs: RNA binding proteins; RF: Random forest; RISC: RNA-induced silencing complex; ROC: Receiver operating characteristic; SVM: Support vector machine

Acknowledgments

We thanks for the three anonymous reviewers's comments on this study.

Funding

This work was supported by Natural Science Foundation of China (No. 61671288, 91530321, 61603161), the Science and Technology Commission of Shanghai Municipality (No. 16JC1404300), and Fellowship from Faculty of Health and Medical Sciences, University of Copenhagen.

Availability of data and materials

The datasets and python code supporting the findings of this study are available at <https://github.com/xypan1232/iDeep> or <http://www.csbio.sjtu.edu.cn/bioinf/iDeep>.

Authors' contributions

XP and HBS designed the study and drafted the manuscript, XP did the bioinformatics analysis. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 15 November 2016 Accepted: 23 February 2017

Published online: 28 February 2017

References

- Ferrè F, Colantoni A, Helmer-Citterich M. Revealing protein-lncRNA interaction. *Brief Bioinform*. 2015;17:106–16. doi:10.1093/bib/bbv031.
- Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009;136:215–33. doi:10.1016/j.cell.2009.01.002.
- Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, Talukder S, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*. 2009;27:667–70. doi:10.1038/nbt.1550.
- Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of the RNA recognition protein and microRNA target sites by PAR-CLIP. *Cell*. 2010;141:129–41. doi:10.1016/j.cell.2010.03.009.
- Stražar M, Žitnik M, Zupan B, Ule J, Curk T. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*. 2016;32:1527–35. doi:10.1093/bioinformatics/btw003.
- Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*. 2014;15:R17. doi:10.1186/gb-2014-15-1-r17.

7. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform.* 2016;17:88–105. doi:10.1093/bib/bbv023.
8. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.* 2015;33:831–8. doi:10.1038/nbt.3300.
9. Pan X, Zhu L, Fan YX, Yan J. Predicting protein-RNA interaction amino acids using random forest based on submodularity subset selection. *Comput Biol Chem.* 2014;53:324–30. doi:10.1016/j.compbiolchem.2014.11.002.
10. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics.* 2006;22:e141–9.
11. Leibovich L, Paz I, Yakhini Z, Mandel-Gutfreund Y. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Res.* 2013;41:W174–9. doi:10.1093/nar/gkt407.
12. Livi CM, Blanzieri E. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. *BMC Bioinforma.* 2014;15:123. doi:10.1186/1471-2105-15-123.
13. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics.* 2004;20:477–86.
14. Kloft M, Brefeld U, Sonnenburg S, Zien A. Lp-norm multiple kernel learning. *J Mach Learn Res.* 2011;12:953–97.
15. Pan X, Xiong K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. *Mol Biosyst.* 2015;11:2219–26. doi:10.1039/c5mb000214a.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436–44.
17. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science.* 2006;313:504–7.
18. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE.* 1998;86:2278–324.
19. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods.* 2015;12:931–4. doi:10.1038/nmeth.3547.
20. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 2016;26:990–9. doi:10.1101/gr.200535.115.
21. LeCun Y, et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* 1989;1:541–51.
22. Zhang C, Yan J, Li C, Rui X, Liu L, Bie R. On Estimating Air Pollution from Photos Using Convolutional Neural Network. *New York: ACM Multimedia (ACM-MM16);* 2016. p. 297–301.
23. Fischer A, Igel C. An Introduction to Restricted Boltzmann Machines. *Lect Notes Comput Sci.* 2012;7441:14–36.
24. Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, Zeng J. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* 2015;44:e32. doi:10.1093/nar/gkv1025.
25. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics.* 2015;31:761–3. doi:10.1093/bioinformatics/btu703.
26. Pan X, Fan YX, Yan J, Shen HB. IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics.* 2016;17:582. doi:10.1186/s12864-016-2931-8.
27. Srivastava N, Salakhutdinov RR. Multimodal learning with deep boltzmann machines. *J Mach Learn Res.* 2014;15:2949–2980.
28. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal Deep Learning. *IEEE Int Conf Mach Learn.* 2011;28:689–96.
29. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol.* 2010;6:e1000832. doi:10.1371/journal.pcbi.1000832.
30. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 2012;40:9379–91. doi:10.1093/nar/gks725.
31. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics.* 2007;23:1495–502.
32. Zdunek R, Cichocki A. Non-negative matrix factorization with quasi-newton optimization. *Artif Intell Soft Comput.* 2006;87:870–9.
33. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA.* 2010;16:1096–107. doi:10.1261/rna.2017210.
34. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature.* 2013;499:172–7. doi:10.1038/nature12311.
35. Pan XY, Tian Y, Huang Y, Shen HB. Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. *Genomics.* 2010;97:257–64. doi:10.1016/j.ygeno.2011.03.001.
36. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome Biol.* 2007;8:R24.
37. Sephton CF, Cenik C, Kucukural A, Dammer EB, Cenik B, Han Y, Dewey CM, Roth FP, Herz J, Peng J, Moore MJ, Yu G. Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *J Biol Chem.* 2011;286:1204–15.
38. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* 2014;15:1929–58.
39. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *Proceedings of The 32nd International Conference on Machine Learning.* vol. 32; 2015. p. 448–56.
40. Lorenz R, Bernhart SH, Hoener zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. *Algorithm Mol Biol.* 2011;6:26.
41. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. *Genome Res.* 2004;14(6):1188–90.
42. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning.* Haifa: Omnipress; 2010. p. 807–814.
43. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 2016;44:e107. doi:10.1093/nar/gkv226.
44. Andrychowicz M, Denil M, Gomez S, Hoffman MW, Pfau D, et al. Learning to learn by gradient descent by gradient descent. 2016. arXiv:1606.04474 [cs.NE].
45. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. *Nucleic Acids Res.* 2016;44:710–6. doi:10.1093/nar/gkv1157.
46. Svetlichnyy D, Imrichova H, Fiers M, Kalender Atak Z, Aerts S. Identification of High-Impact cis-Regulatory Mutations Using Transcription Factor Specific Random Forest Models. *PLoS Comput Biol.* 2015;11:e1004590. doi:10.1371/journal.pcbi.1004590.
47. Frith MC, Li MC, Weng Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* 2003;31:3666–8.
48. Smolensky P. Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. Cambridge: MIT Press; 1986, p. 194–281.
49. Hinton GE. A practical guide to training restricted Boltzmann machines. *Momentum.* 2010;9:926.
50. Tieleman T, Hinton GE. Lecture 6.5 - rmsprop: Divide the gradient by a run-ning average of its recent magnitude. COURSE: Neural Netw Mach Learn. 2012;4:2.
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825–30.